Cricket IPL - Indian Premier League

Aura Petrucci: 626712
Federica Altavilla: 626718

Progetto di Analisi di dati

Sommario

Il presente lavoro nasce con l'obiettivo di svolgere un'analisi riguardante i dati dell'Indian Premier League, con un maggior riguardo sulle variabili ritenute più significative e le loro eventuali correlazioni. L'analisi si compone delle seguenti fasi: preparazione del dataset, data understanding e outlier detection.

---

1 Introduzione al dataset

I dati analizzati sono suddivisi in due dataset distinti (deliveries.csv e matches.csv), entrambi relativi ad osservazioni riguardanti 636 partite di cricket della Premier League indiana giocate dal 2008 al 2016. In particolare il primo fa riferimento ai dati per ogni palla dei match giocati (battitore, punti segnati, espulsioni ecc.), mentre nel secondo si fa riferimento ai dati relativi ad ogni partita, come: risultati, arbitri, squadre, città dove viene giocata la partita, ecc.

2 Preparazione del dataset

I due dataset si compongono di variabili di diverso tipo. *Deliveries* presenta in totale 21 features, delle quali 13 sono numeriche di tipo intero, mentre le restanti di tipo categorico. Per quel che riguarda il dataset *Matches*, avente 18 variabili, 6 di esse sono di tipo numerico, mentre le restanti sono di tipo categorico. Le tabelle sottostanti (tab. 1 e 2) sono state realizzate per avere una migliore comprensione della suddivisione delle varie features nei due dataset.

Variabili categoriche	*match_id*, *batting_team*, *bowling_team*, *batsman*, *non_strike*, *bowler*, *player_dismissed*, *dismissal_kind*, *fielder*
Variabili numeriche discrete	*inning*, *over*, *ball*, *is_super_over*, *wide_runs*, *bye_runs*, *legbye_runs*, *no_ball_runs*, *penalty_runs*, *batsman_runs*, *extra_runs*, *total_runs*

Tabella 1: Variabili del dataset *deliveries*

Variabili categoriche	*city*, *date*, *team1*, *team2*, *toss_winner*, *toss_decision*, *result*, *winner*, *venue*,*player_of_match*, *umpire1*, *umpire2*
Variabili numeriche discrete	*id*, *season*, *dl_applied*, *win_by_runs*, *win_by_wickets*
Variabili numeriche continue	*umpire3*

Tabella 2: Variabili del dataset *matches*

In un primo momento, vista la natura dei dati presenti all'interno dei datasets e la presenza di due features equivalenti in entrambe (*id* e *match_id*), si è pensato di procedere attraverso la fusione dei due in un unico dataset per una migliore gestione delle informazioni. Il metodo **.merge()**, ne avrebbe realizzato l'unione seppur le variabili in comune presentano delle denominazioni differenti. Tuttavia, si è convenuto che una tale fusione non avrebbe giovato all'analisi dei dati, per tal motivo si è preferito lavorare sui due dataset separatamente.

2.1 Gestione dei valori mancanti

I datasets analizzati presentano numerosi valori mancanti, per i quali diverse soluzioni sono possibili. In base a quanti *missing values* sono stati reperiti in ogni variabile, si è deciso di operare:

- eliminando l'intera feature;
- sostituendo i valori mancanti;

Variabili	Missing Values	% Missing Values
umpire3	636/636	100%
city	7/636	1.1%
winner	3/636	0.5%
player_of_match	3/636	0.5%
umpire1	1/636	0.2%
umpire2	1/636	0.2%
fielder	145091/150460	96.4%
player_dismissed	143022/150460	95.1%
dismissal_kind	143022/150460	95.1%

Tabella 3: Valori mancanti e la loro percentuale

Inizialmente, è stato calcolato il numero e la percentuale di *missing values* presenti nelle variabili. Come riportato nella tabella 3, si può osservare che le features *umpire3*, *fielder*, *player_dismissed* e *dismissal_kind* presentano una percentuale maggiore del 95% di *missing values*.

In un primo momento, si è pensato di mantenere le features *player_dismissed* e *dismissal_kind*, in quanto mostrano informazioni potenzialmente interessanti considerando che indicano rispettivamente:

- i giocatori squalificati dalla partita;
- la tipologia di squalifica applicata.

Inoltre, tali valori mancanti, non indicano necessariamente un dato assente o incompleto, bensì rappresentano l'effettiva mancanza di giocatori squalificati durante il *match*. In ogni caso, si è infine optato per eliminare definitivamente tali *features* in quanto la percentuale di valori mancanti era vicina al 100%. I restanti valori mancanti presentano invece percentuali molto più basse, per cui, per le variabili *winner*, *player_of_match*, *umpire1* e *umpire2*, si è deciso di eliminare le righe contenenti questi valori nulli, attraverso il metodo **.dropna()**.

Per concludere, la variabile *city*, presentando solo l'1.1 % di *missing values*, è stata sostituita. Per realizzare questa operazione sono state individuate le righe contenenti tali dati assenti, e osservando in particolare la feature *venue*, i.e., stadio nel quale la partita si è disputata, si è potuto risalire alla città. In questo caso, lo stadio in questione era il *Dubai International Criket Stadium*, che ha appunto sede nella città di Dubai.

2.2 Data refinement: la variabile *dl_applied*

Uno studio più approfondito è stato realizzato per la variabile *dl_applied*: essa sta ad indicare se in una partita è stata applicata la formula Duckworth-Lewis, metodo matematico utilizzato per calcolare i punteggi conclusivi da assegnare alle squadre in casi di partite fortemente influenzate dalle condizioni atmosferiche. Nella fase di data understanding, osservando le statistiche della feature in questione, si è riscontrato che essa assume solamente valori pari a 0 e 1, che indicano l'applicazione o meno di tale metodo durante un match. Per tal motivo si è proceduto alla conversione del tipo della variabile: da integer, attraverso il metodo **.astype('bool')** essa è stata trasformata in boolean.

3 Data Understanding

Nella fase di data understanding, attraverso il metodo **.describe()**, si sono analizzate le statistiche di ciascuna variabile numerica di entrambi i dataset; alcune delle più interessanti sono riportate nella tabella sottostante.

	noball_runs	extra_runs	total_runs	win_by_runs	win_by_wickets
mean	0.00	0.07	1.29	13.84	3.29
std	0.07	0.35	1.58	23.71	3.38
min	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00	0.00
50%	0.00	0.00	1.00	0.00	3.00
70%	0.00	0.00	1.00	20.00	6.00
max	5.00	7.00	7.00	146.00	10.00

Tabella 4: Stime statistiche di alcune features

Come è possibile osservare, per la variabile *win_by_runs* si ha una media di 13.84 ed un valore massimo di 146.00, tali valori sono quindi il sentore di una distribuzione sbilanciata delle vincite per punteggio. Per tal motivo, si è deciso di visualizzare, attraverso un istogramma, la distribuzione dei valori di tale variabile; dall'istogramma che ne risulta (fig. 1) si nota come nella maggior parte delle partite si abbiano poche *win_by_runs*.

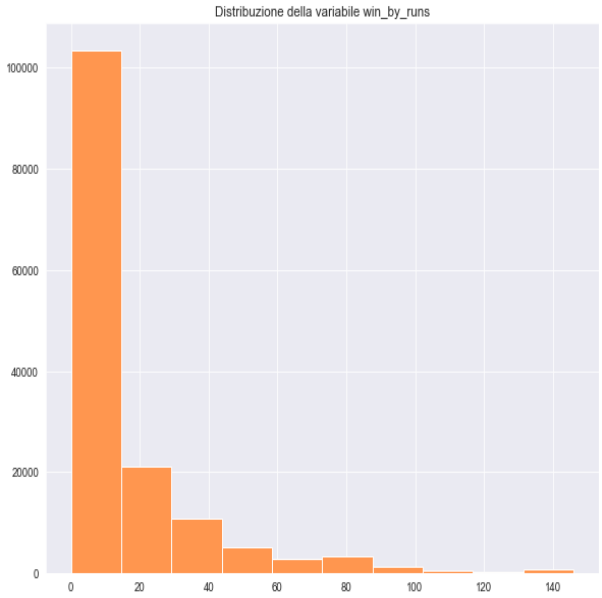

Figura 1: Distribuzione delle vittorie per punteggio

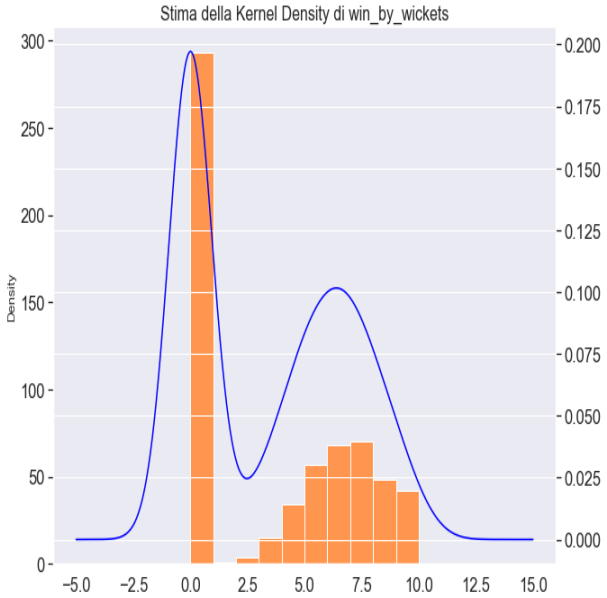

Figura 2: Kernel Density estimation di *win_by_wickets*

Trattandosi quindi di una distribuzione non normale, si è deciso di calcolare l'indice di *skewness* della variabile, che conferma l'andamento asimmetrico di essa con un indice di *Kurtosis* estremamente elevato (cfr. tabella 5). Oltre a questa, sono stati riportati gli indici di asimmetria e di *Kurtosis* delle features *win_by_wickets*, *total_runs* e *batsman_runs*, per le quali rimarchiamo:

- una distribuzione leggermente asimmetrica per *win_by_wickets* (fig. 2)
- un'asimmetria molto marcata sia per *total_runs* (avente un'asimmetria di 1.56) che per *batsman_runs* (1.59).

	skewness	**kurtosis**
win_by_runs	2.50	7.33
win_by_wickets	0.25	-1.53
total_runs	1.56	1.69
batsman_runs	1.59	1.60

Tabella 5: Indici di Kurtosis e Skewness di alcune variabili

Il dataset *deliveries* è stato esplorato più nel dettaglio attraverso la visualizzazione di alcune features considerate interessanti. Si è proseguito osservando la differenza nella distribuzione tra i *total_runs* ed altre variabili ovvero *wide_runs*, *bye_runs*, *noball_runs* e *legbye_runs*, le quali corrispondono a eventuali penalità che una delle due squadre può subire, a favore dell'altra che guadagna punti extra. Di queste, solamente la prima è stata riportata (fig. 3), in quanto presentavano tutte una distribuzione simile, dove il numero di runs risulta sempre moderatamente basso. Nella figura 4 si può osservare invece il confronto tra le distribuzioni di *total_runs* e *batsman_runs* (indicante i punti segnati dal battitore), che presentano tra loro valori più analoghi rispetto alle precedenti.

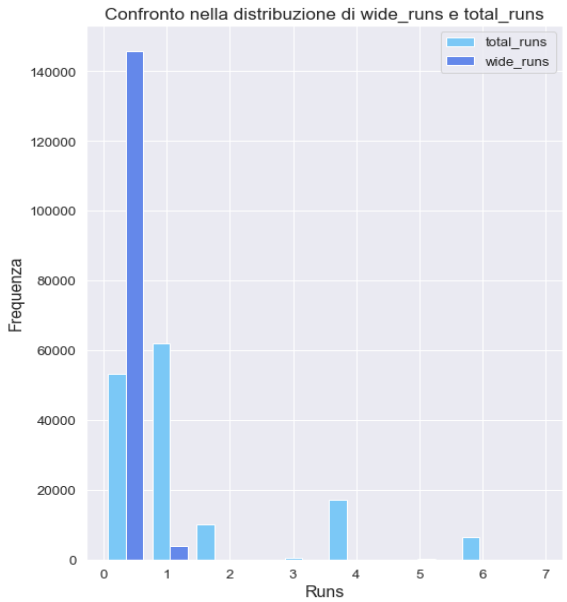

Figura 3: Confronto nella distribuzione di *wide_runs* e *total_runs*

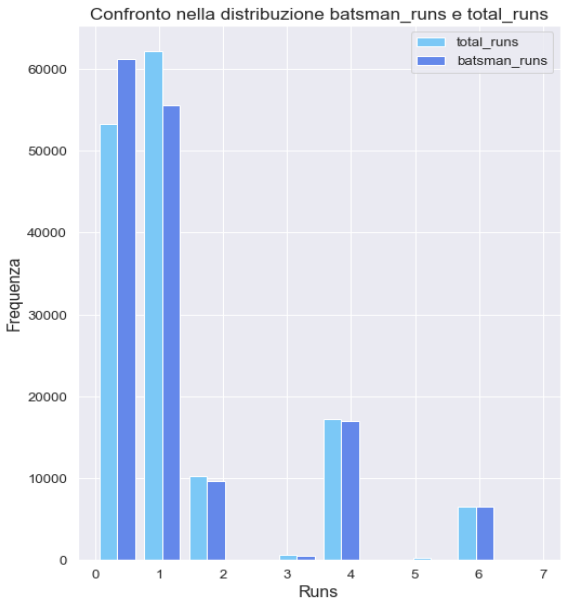

Figura 4: Confronto nella distribuzione di *batsman_runs* e *total_runs*

In seguito l'analisi è proseguita analizzando la distribuzione di alcune variabili del dataset *matches*; l'istogramma in figura 5, ad esempio, rappresenta la distribuzione del numero di partite giocate in ogni stadio. Come si può notare, gli stadi con il maggior numero di match giocati sono: *M. Chinnaswamy Stadium*, *Eden Gardens* e *Wankhede Stadium*. Per quel che riguarda le differenti stagioni, il 2011, 2012 e 2013 risultano essere gli anni nei quali si sono giocate più partite, come è possibile osservare dal grafico in figura 6.

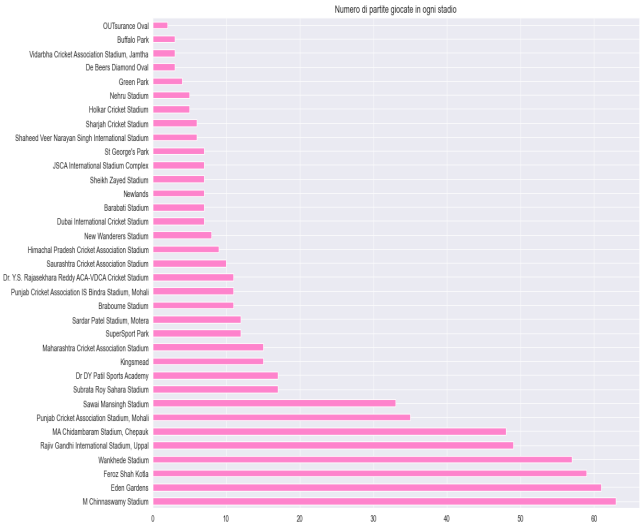

Figura 5: Partite giocate per stadio

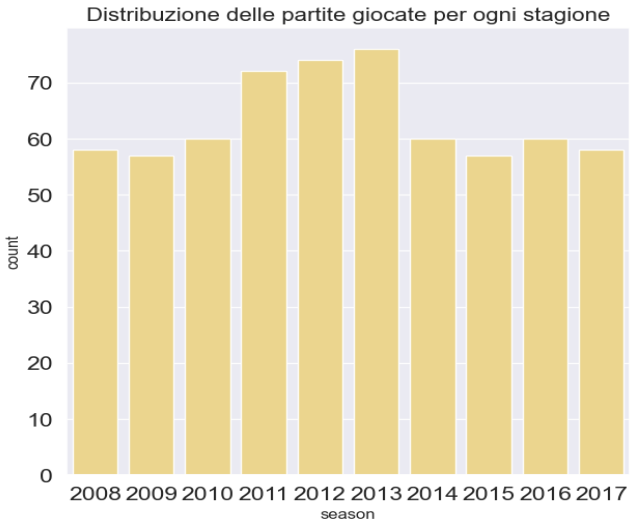

Figura 6: Partite giocate per stagione

Durante lo studio delle variabili ci si è avveduti della presenza di un errore di tipo sintattico nella feature *winner*, dove apparivano come teams distinti *"Rising Pune Supergiants"* e *"Rising Pune Supergiant"*; a seguito di una ricerca si è convenuto che esse venivano considerate come due squadre distinte seppur riferendosi alla stessa. Per tal motivo, al fine di evitare un'analisi errata e una distribuzione distorta delle features, le celle contenenti la squadra *"Rising Pune Supergiants"* sono state sostituite con "*Rising Pune Supergiant"* attraverso il metodo **.replace()**. A seguito di ciò, grazie ad una visualizzazione attraverso un piechart della variabile *"winner"* (fig. 7), si evince che la squadra più forte dell'Indian Premier League sia *Mumbai Indians*. Inoltre, attraverso l'utilizzo del metodo **.value_counts()**, ne è stato calcolato l'esatto numero di partite vinte durante l'IPL, pari a 92.

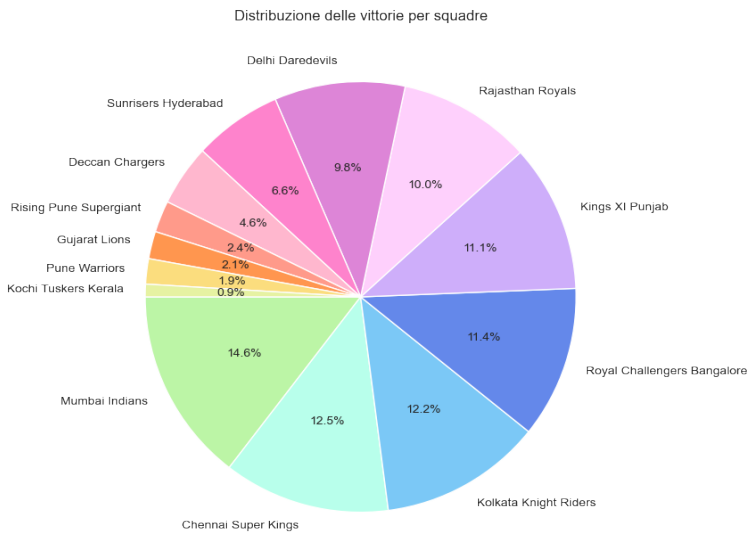

Figura 7: Distribuzione delle partite vinte da ogni squadra

Per un'analisi più approfondita riguardo le squadre vincenti, si è deciso di visualizzare la distribuzione delle vittorie per squadra in base alla stagione. La figura 8 mostra dei risultati interessanti, ad esempio i *Chennai Super Kings* hanno vinto diverse partite dal 2008 al 2015, a differenza degli ultimi due anni del campionato. Ugualmente per i *Deccan Chargers* le vittorie si concentrano tra il 2008 e il 2012.
Attraverso il **.value_counts()** della variabile *player_of_match* è stato possibile calcolare quante volte ogni giocatore è stato nominato miglior giocatore di una partita. Con tali risultati si è in seguito implementato un istogramma (fig. 9) che mostra quali sono i giocatori migliori di tutte le stagioni, ovvero *CH Gayle e YK Patahan*, in quanto sono stati nominati più di 15 volte come "player of the match".

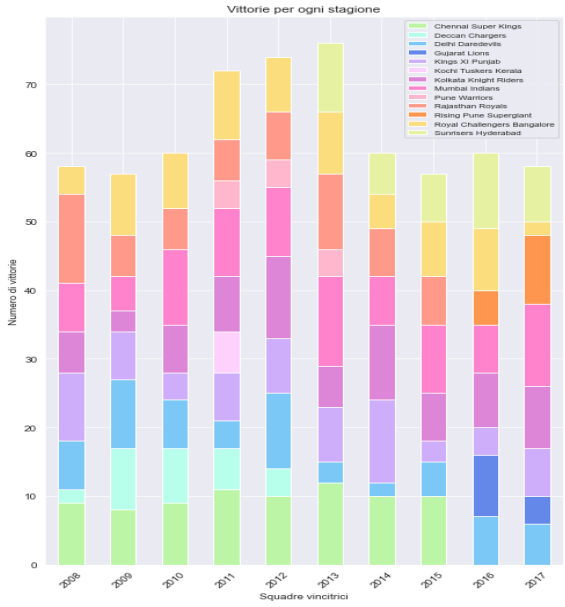

Figura 8: Distribuzione delle vittorie di ogni squadra per stagione

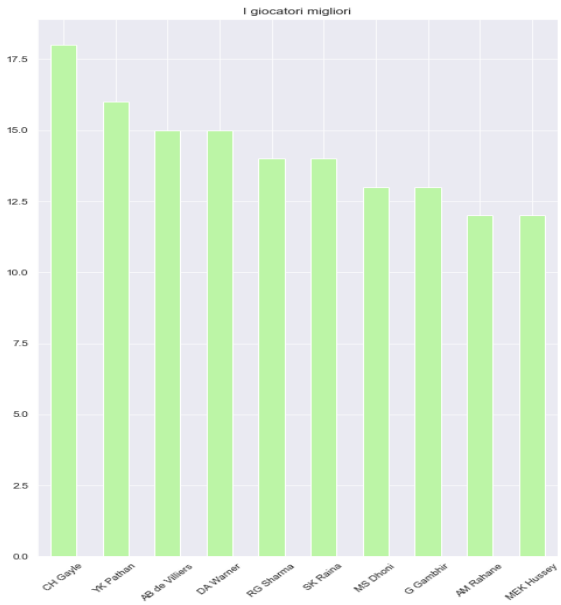

Figura 9: I migliori giocatori dal 2008 al 2016

Un altro aspetto approfondito riguarda il confronto tra le vittorie e le sconfitte di ogni squadra durante l'Indian Premiere League. Dalla figura 10 si possono reperire diverse informazioni: un esempio interessante concerne le prime tre squadre riportate (*Kochi Tuskers Kerala*, *Pune Warriors*, *Gujarat Lions*), che presentano un numero totale di partite giocate sostanzialmente inferiore rispetto alle restanti; il motivo di ciò può essere spiegato osservando la figura 8, attraverso la quale comprendiamo che tali squadre non hanno partecipato a tutte le stagioni del campionato.

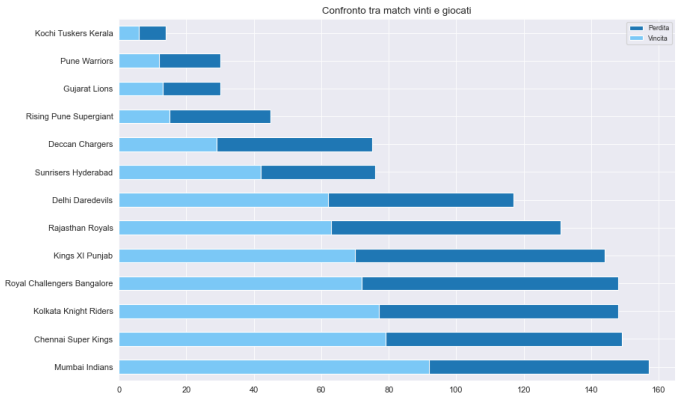

Figura 10: Confronto tra partite vinte e perse da ogni squadra

In seguito è stato calcolato il numero di *runs* dei battitori (*batsman_runs*) e la loro squadra di appartenenza (fig. 11). L'analisi si è limitata ai 10 battitori che hanno registrato il numero maggiore di punti; come si può osservare, tre di questi appartengono alla *Royal Challengers Bangalore*, la quale inoltre risulta essere una delle squadre con la percentuale di vittorie più alta (cfr. fig.7).

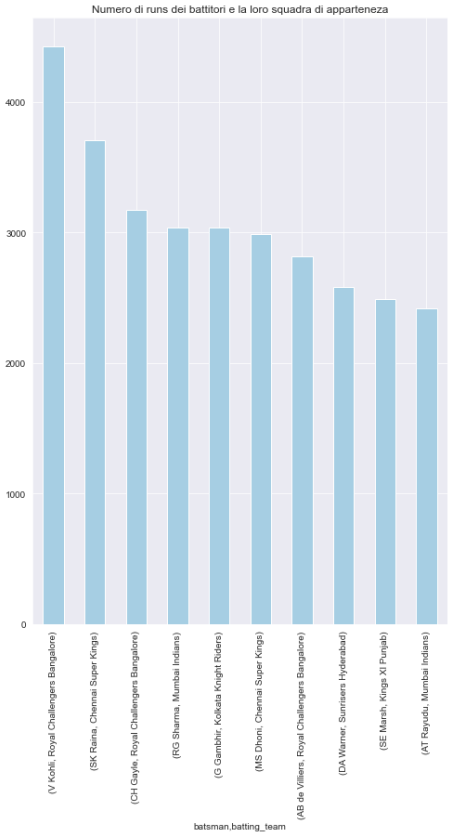

Figura 11: Runs dei battitori e la loro squadra di appartenenza

L'analisi è proseguita osservando se ci fosse un possibile legame, tra le vittorie delle squadre e la precedente "vincita" al lancio della moneta (*toss_winner*). Per fare ciò è stata verificata la possibile corrispondenza tra i record delle variabili *toss_winner* e *winner*:

$$\textbf{probabilita\_vincita} = \textbf{matches['toss\_winner']} == \textbf{matches['winner']}$$

I risultati ottenuti non mostrano però una forte relazione causale tra le due: la figura 12 mostra infatti che la corrispondenza tra *winner* e *toss_winner* non è tanto più rimarcata rispetto alla non corrispondenza delle stesse.
Per tal motivo si è deciso di approfondire lo studio realizzando una **crosstab** che mostra la scelta presa (*toss_decision*) da ogni squadra vincente del match (*winner*) e del lancio della moneta (*toss_winner*). Dai risultati ottenuti (fig. 13) si rimarca che nei 201 casi in cui la scelta è ricaduta su *field* la squadra ha vinto il match; tale decisione potrebbe quindi aver influito positivamente sull'esito finale delle partite.

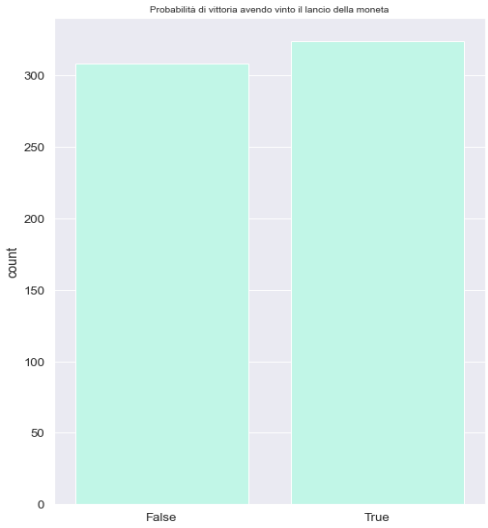

Figura 12: Probabilità di vincita dopo il lancio della moneta

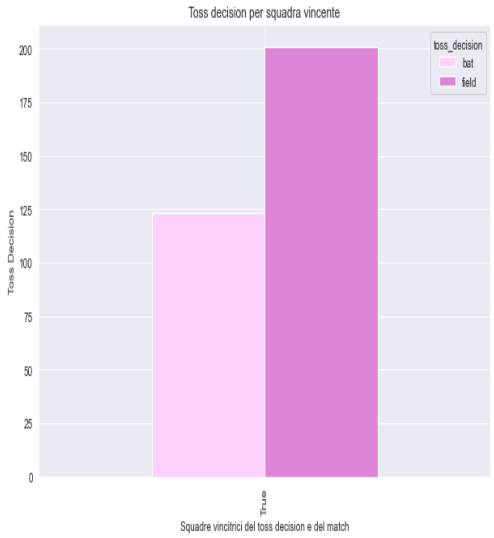

Figura 13: Toss decision per squadra vincente

Per calcolare la quantità di punti totalizzati durante ogni stagione (*total_runs)* si è proseguito attraverso un **.merge()** tra il dataset *deliveries* e due variabili del dataset *matches*; tale unione è per l'appunto avvenuta attraverso le due variabili in comune presenti in entrambi, ovvero *id e match_id*. In tal modo è stato possibile plottare un grafico lineare (fig. 14) dove si può osservare che le stagioni in cui sono stati realizzati più punti totali sono il 2012 e il 2013, ovvero gli anni in cui sono state giocate più partite (cfr. fig. 6).

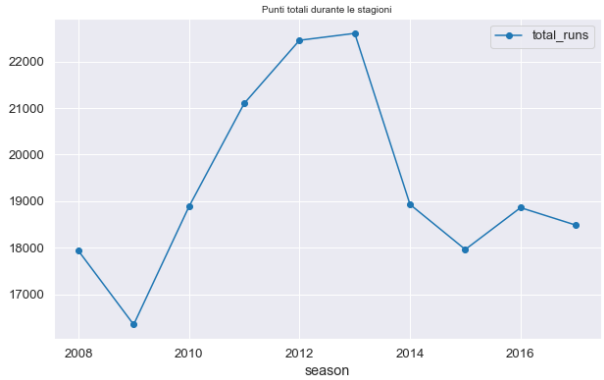

Figura 14: Numero di *total_runs* totale per ogni stagione

3.1 Focus sulle variabili *dl_applied* e *season*

La variabile *dl_applied* ha suscitato particolare interesse e per questo si è deciso di verificare l'applicazione della formula a cui essa si riferisce nell'arco delle diverse stagioni. Per fare ciò, è stata realizzata una **crosstab** attraverso la quale è emerso che solo in occasione di 16 partite il metodo *Duckworth-Lewis* è stato implementato. Più nel dettaglio, è stato maggiormente impiegato durante le stagioni 2009, 2011 e 2016. Di conseguenza si è voluto osservare se, in quelle tre annate, la sua applicazione fosse avvenuta nello stesso periodo dell'anno ed eventualmente in corrispondenza della stagione piovosa in India. Attraverso un'osservazione più attenta, si è evinto che, in tutti e tre i casi, tali partite si sono disputate nei mesi di aprile e maggio. Ciò potrebbe indicare delle condizioni atmosferiche particolarmente avverse in quel periodo dell'anno: quest'informazione risulta peculiare e rara in quanto la stagione dei monsoni in India non inizia prima del mese di giugno.

4 Correlazione tra le variabili

Lo studio è proseguito osservando eventuali correlazioni che intercorrono tra le features dei dataset analizzati. Innanzitutto, attraverso il metodo **.corr()**, che calcola di default il coefficiente di correlazione lineare di Pearson, è stato possibile ottenere una panoramica generale della correlazione tra ogni coppia di variabili del dataset *deliveries*.

Attraverso la heatmap qui riportata (fig. 15) è possibile avere una visualizzazione grafica di tale correlazioni. La maggior parte di esse risultano scorrelate tra loro, con due sole eccezioni:

- *wide_runs* e *extra_runs* con una correlazione di 0.72;
- *total_runs* e *batsman_runs* con un coefficente pari a 0.98.

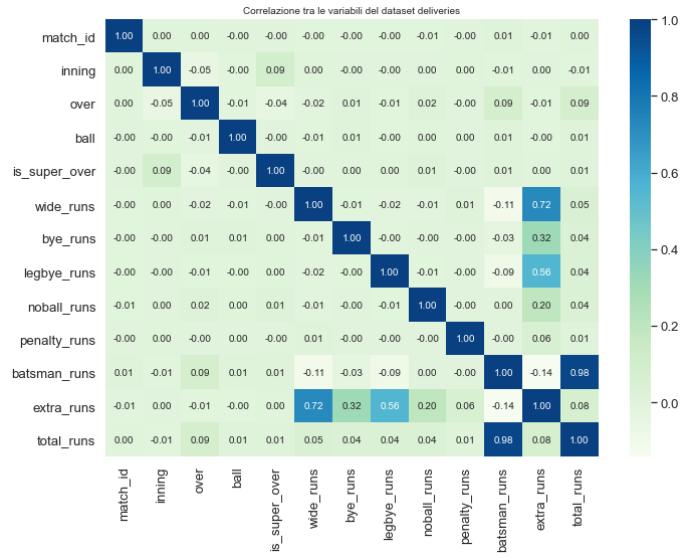

Figura 15: Heatmap delle correlazione tra le variabili del dataset *Deliveries*

Si è proseguito calcolando inoltre il Coefficiente di Correlazione per ranghi di Spearman e Kendall, per osservare eventuali cambiamenti. Tuttavia, anche in questi casi le correlazioni interessanti si hanno tra *wide_runs* e *extra_runs*, e tra *total_runs* e *batsman_runs*, con un coefficiente (in entrambi i casi) pari a 0.73 nella prima e 0.93 nella seconda coppia di valori. Ciò dimostra che tali coppie di variabili siano correlate positivamente, ovvero all'aumentare di una aumenta di conseguenza anche l'altra; quindi si può evincere che nel caso dei punti totali (*total_runs*), la maggior parte di questi sono stati realizzati dal battitore (*batsman_runs*) ed inoltre che i punti extra (*extra_runs*) siano principalmente guadagnati attraverso le *wide_runs*, ovvero punteggi per tiri lunghi. In effetti la palla larga (wide ball) è uno dei metodi più comuni per ottenere corse extra nel cricket. Vista la forte correlazione tra le variabili *total_runs e batsman_runs*, è stata plottata la regressione lineare tra le due e riportata nella figura qui presente (fig. 16); ciò è stato possibile definendo le variabili x e y e il dataframe deliveries sul quale applicare la regressione. Si può osservare come si possa prevedere il valore della variabile *batsman_runs* in base al valore di *total_runs*, i.e., della variabile indipendente.

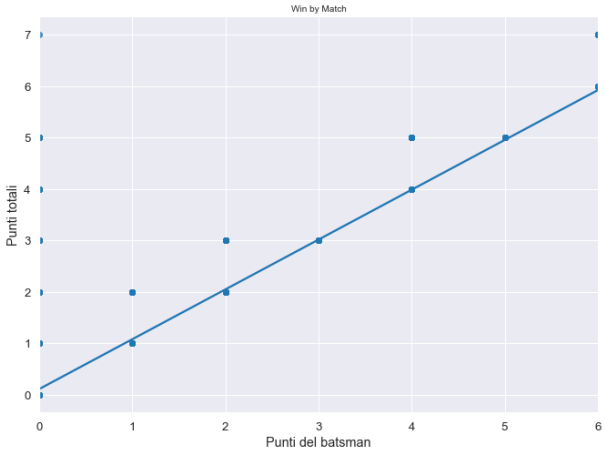

Figura 16: Regressione Lineare della variabile *batsman_runs*

A ulteriore conferma sono stati calcolati due scatter plot (fig. 17 e 18) per tali coppie, ottenendo come risultati, in entrambi i casi, delle rappresentazioni dove si visualizza chiaramente una relazione di tipo lineare tra le due variabili.

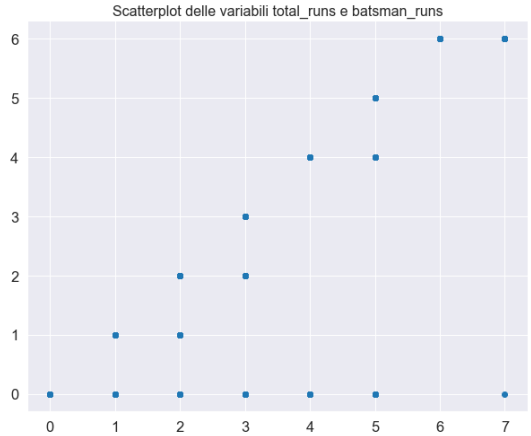

Figura 17: Scatterplot di total_runs e batsman_runs

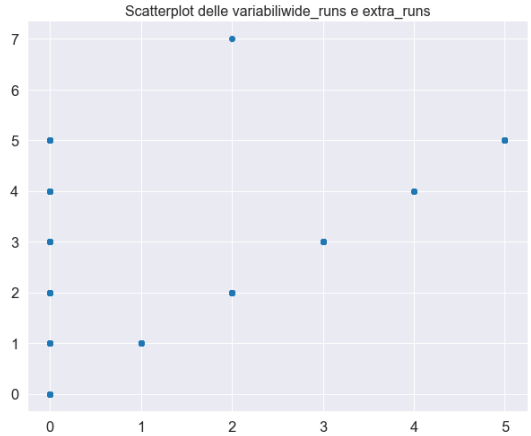

Figura 18: Scatterplot di wide_runs e extra_runs

Per quel che riguarda il dataset *matches*, anche in questo caso è stata calcolata la correlazione di Pearson, Spearman e Kendall tra ogni coppia di variabili numeriche, senza ottenere però dei risultati interessanti, ad eccezione di una correlazione di tipo negativo tra *win_by_runs* e *win_by_wickets*. In questo caso il coefficiente Pearson equivale a -0.58, quello di Spearman è pari a -0.84 mentre quello di Kendall è uguale a -0.67. Tale correlazione risulta ovvia in quanto è scontato che la vittoria di un determinato tipo (che essa sia *win_by_wickets* o *win_by_runs*) escluda l'altra.

5 Outlier Detection

Come ultima fase dello studio è stata realizzata un'analisi degli outliers dei due datasets. Per fare ciò, per il dataset *matches*, sono state prese in considerazione le variabili: *win_by_runs e win_by_wickets*; di queste sono stati calcolati in un primo momento il primo e il terzo quartile. Da tali risultati è stato calcolato quindi il range interquartile ed infine l'upperfence ed il lowerfence per ogni variabile, ovvero i limiti massimi e minimi al di fuori dei quali i valori che vi ricadono sono considerati outliers, in quanto anomali rispetto all'andamento generico della feature. Attraverso i boxplot si ottiene una visualizzazione grafica di tali valori; in questo report è stata riportata solamente la rappresentazione degli outliers per *win_by_runs*, in quanto risulta essere la sola rilevante. In effetti per *win_by_wickets* non si hanno valori anomali, ovvero tutti i suoi records presentano dei punteggi più o meno analoghi tra di loro; nella figura 19 si può osservare invece una situazione opposta, nella quale il boxplot individua la presenza di outliers, localizzati principalmente al di sopra dell'*upperfence*, i.e., corrispondono a valori molto più elevati rispetto alla media. Tuttavia, si è optato per mantenere nel dataset questi record, in quanto rappresentano semplicemente una performance sportiva notevolmente superiore rispetto alla norma.

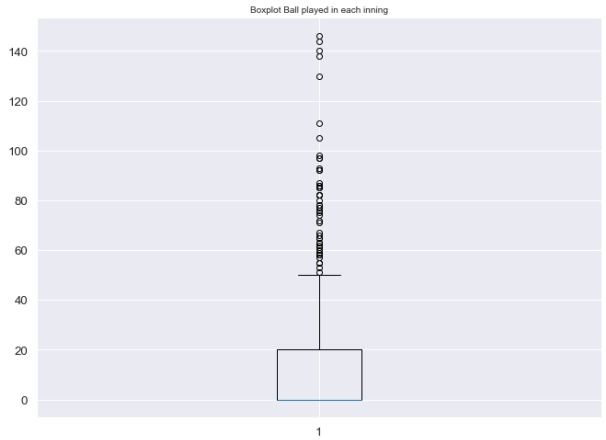

Figura 19: Visualizzazione tramite boxplot degli outliers della variabile *win_by_runs*

In un secondo momento, si è deciso di analizzare anche gli outliers della variabile *player_of_match*; per tale analisi è stato utilizzato il metodo **.value_counts()** per verificare quante volte apparissero i nomi dei giocatori considerati come migliori. Per l'*outliers detection* sono stati quindi utilizzati i valori numerici corrispondenti alle volte nelle quali tali giocatori sono stati nominati *player_of_match*.

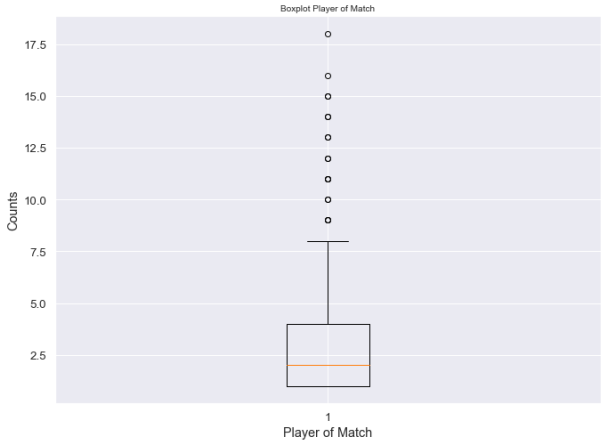

Figura 20: Visualizzazione tramite boxplot degli outliers della variabile *player_of_match*

Il boxplot nella figura 20 mostra la distribuzione della variabile in questione, la quale indica la presenza di outliers solamente nella parte superiore rispetto all'upperfence, ciò indica quindi che i valori anomali che vengono trovati sono rappresentati unicamente dai numeri più elevati, ovvero dai giocatori che sono stati più volte nominati come migliori del match; per tal motivo una gestione di tali outliers non avrebbe senso ai fini del nostro studio, essi sono stati quindi mantenuti.

L'outlier detection è stata realizzata anche per tutte le variabili di tipo numerico presenti nel dataset deliveries; i risultati più significativi sono riportati qui di seguito. Per le features *ball* e *over* non sono stati reperiti dati anomali, vale a dire che i valori rientrano nel range interquartile, per il primo tra 2 e 5, mentre per il secondo tra 5 e 15 (fig.21). Al contrario, la variabile *batsman_runs* presenta numerosi outliers (fig.22) indicanti un numero di punti segnati dai battitori, durante determinate partite, molto più elevato rispetto alla norma.

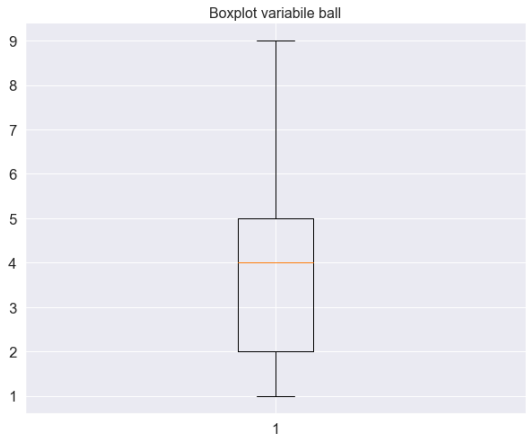

Figura 21: Boxplot della variabile *over*

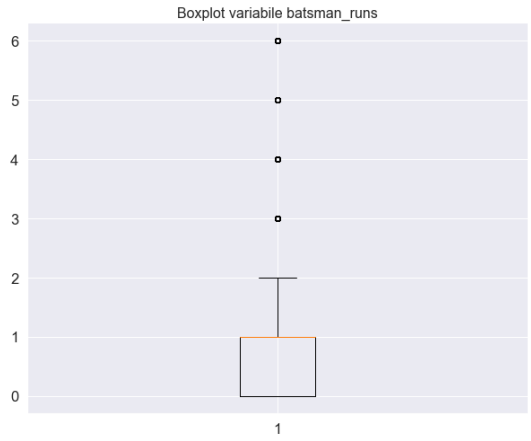

Figura 22: Boxplot della variabile *batsman_runs*

Anche in questo caso, si è optato per mantenere e non modificare gli outliers identificati, in quanto non riguardano errori di misurazioni tecniche, ma performance migliori rispetto alla media.

6 Conclusioni

Con il presente lavoro sono state analizzate le diverse features dei due datasets, ottenendo informazioni interessanti riguardo le partite giocate durante l'Indian Premier League, come: partite giocate per stagione, migliori giocatori, stadi in cui si sono disputati più match etc. Inoltre è stata svolta un'analisi per eventuali correlazioni tra coppie di variabili ed infine un calcolo di eventuali outliers i quali però, per la loro natura, non sono stati trattati o eliminati. Lo studio si è incentrato sulla descrizione delle caratteristiche delle differenti variabili, ma l'analisi potrebbe essere approfondita attraverso l'implementazione di algoritmi di data mining come clustering e classificazione, i quali possono aiutare in una predizione futura riguardo le successive stagioni dell'IPL.